Towards Practical Automated Human Action Recognition

A Thesis Submitted for the Degree of Doctor of Philosophy

By

Zia Moghaddam

in

School of Computing and Communications

Faculty of Engineering and Information Technology

UNIVERSITY OF TECHNOLOGY, SYDNEY (UTS)

AUSTRALIA

March 2012

CERTIFICATE

Date: **1st March 2012**

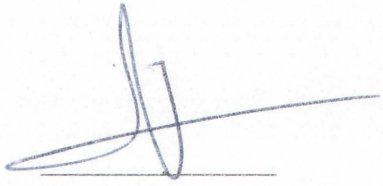
Author: **Zia Moghaddam**

Title: **Towards Practical Automated Human Action Recognition**

Degree: **Ph.D.**

I certify that this thesis has not already been submitted for any degree and is not being submitted as part of candidature for any other degree.

I also certify that the thesis has been written by me and that any help that I have received in preparing this thesis, and all sources used, have been acknowledged in this thesis.



Signature of Author

This thesis is dedicated to my wife, Hamideh,

and my two sons, Mohammad Hossein and Mohammad Sadra

Acknowledgments

First of all, I am very appreciative of my supervisor, Prof. Massimo Piccardi, for his complete guidance and direction through my PhD, for being always willing to answer questions, for his understanding of my family situation, and for his full support and encouragement throughout all years of my study.

I also thank my friends and colleagues at the School and our research group, iNEXT, for their kindness and support.

Furthermore, this author is grateful to SenSen Pty Ltd for partially supporting the work presented in this thesis.

Finally, thanks to my wife, Hamideh, for her tremendous support and encouragement which allowed me to pursue my research.

Author's Publications

The following publications have been produced during the course of this thesis:

Conference papers:

- **Z. Moghaddam** and M. Piccardi, "Deterministic Initialization of Hidden Markov Models for Human Action Recognition", *Proc. of the Digital Image Computing: Techniques and Applications (DICTA)*, 2009, pp. 188-195.
- **Z. Moghaddam** and M. Piccardi, "Histogram-Based Training Initialisation of Hidden Markov Models for Human Action Recognition", *Proc. of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2010, pp. 256-261.
- **Z. Moghaddam** and M. Piccardi, "Human Action Recognition with MPEG-7 Descriptors and Architectures", *Proc. of the 1st ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS)*, 2010, pp. 63-68.
- **Z. Moghaddam** and M. Piccardi, "Robust Density Modelling using the Student's t-distribution for Human Action Recognition", *Proc. of the 18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 3261-3264.
- O.P. Concha, R. Xu, **Z. Moghaddam** and M. Piccardi, "HMM-MIO: An Enhanced Hidden Markov Model for Action Recognition," *Proc. of the 3rd International Workshop on Machine Learning for Vision-based Motion Analysis (MLvMA) in conjunction with CVPR 2011*, pp. 62-69.

Table of Contents

List of Figures

List of Tables

Abstract

Modern video surveillance requires addressing high-level concepts such as humans' actions and activities. Automated human action recognition is an interesting research area, as well as one of the main trends in the automated video surveillance industry. The typical goal of action recognition is that of labelling an image sequence (video) using one out of a set of action labels. In general, it requires the extraction of a feature set from the relevant video, followed by the classification of the extracted features. Despite the many approaches for feature set extraction and classification proposed to date, some barriers for *practical action recognition* still exist. We argue that recognition accuracy, speed, robustness and the required hardware are the main factors to build a practical human action recognition system to be run on a typical PC for a real-time video surveillance application. For example, a computationally-heavy set of measurements may prevent practical implementation on common platforms.

The main focus of this thesis is challenging the main difficulties and proposing solutions towards a practical action recognition system. The main outstanding difficulties that we have challenged in this thesis include 1) initialisation issues with model training; 2) feature sets of limited computational weight suitable for real-time application; 3) model robustness to outliers; and 4) pending issues with the standardisation of software interfaces. In the following, we provide a description of our contributions to the resolution of these issues.

Amongst the different classification approaches for classifying actions, graphical models such as the hidden Markov model (HMM) have been widely exploited by many researchers. Such models include observation probabilities which are generally modelled by mixtures of Gaussian components. When learning an HMM by way of Expectation-Maximisation (EM) algorithms, arbitrary choices must be made for their initial parameters. The initial choices have a major impact on the parameters at convergence and, in turn, on the recognition accuracy. This dependence forces us to repeat training

with different initial parameters until satisfactory cross-validation accuracy is attained. Such a process is overall empirical and time consuming.

We argue that one-off initialisation can offer a better trade-off between training time and accuracy, and as one of the main contributions of this thesis, we propose two methods for deterministic initialisation of the Gaussian components' centres. The first method is a time segmentation-based approach which divides each training sequence into the requested number of clusters (product of the number of HMM states and the number of Gaussian components in each state) in the time domain. Then, clusters' centres are averaged among all the training sequences to compute the initial centre for each Gaussian component. The second approach is a histogram-based approach which tries to initialise the components' centres with the more popular values among the training data in terms of density (similar to mode seeking approaches). The histogram-based approach is performed incrementally, considering each feature at a time. Either centre initialisation approach is followed by dispatching the resulting Gaussian components onto HMM states. The reference component dispatching method exploits the arbitrary order for dispatching. In contrast, we again propose two more intelligent methods based on the effort to put components with closer centres in the same state which can improve the correct recognition rate.

Experiments over three human action video datasets (Weizmann [1], MuHAVi [2] and Hollywood [3]) prove that our proposed deterministic initialisation methods are capable of achieving accuracy above the average of repeated random initialisations (about 1 per cent to 3 per cent in 6 random runs experiment) and comparable to the best. At the same time, one-off deterministic initialisation can save the required training time substantially compared to repeated random initialisations, e.g. up to 83% in the case of 6 runs of random initialisation. The proposed methods are general as they naturally extend to other models where observation densities are conditioned on discrete latent variables, such as dynamic Bayesian networks (DBNs) and switching models.

As another contribution, we propose a simple and computationally lightweight feature set, named *sectorial extreme points*, which requires only 1.6 ms per frame for extraction on a reference PC. We believe a lightweight feature set is more appropriate for the task of action recognition in real-time surveillance applications with the usual requirement of processing 25 frames per second (PAL video

rate). The proposed feature set represents the coordinates of the extreme points in the contour of a subject's foreground mask. The various experiments prove the strength of the proposed feature set in terms of classification accuracy, compared to similar feature sets, such as the star skeleton [4] (by more than 3%) and the well-known projection histograms (up to 7%).

Another main issue in density modelling of the extracted features is the outlier problem. The extraction of human features from videos is often inaccurate and prone to outliers. Such outliers can severely affect density modelling when the Gaussian distribution is used as the model since it is short-tailed and highly sensitive to outliers. Hence, outliers can affect the classification accuracy of the HMM-based action recognition approaches that exploit Gaussian distribution as the base component. In contrast, the Student's *t*-distribution is more robust to outliers thanks to its longer tail and can be exploited for density modelling to improve the recognition rate in the presence of abnormal data. As another main contribution, we present an HMM which uses mixtures of *t*-distributions as observation probabilities and apply it for the recognition task. The conducted experiments over the Weizmann and MuHAVi datasets with various feature sets report a remarkable improvement of up to 9% in classification accuracy by using HMM with mixtures of *t*-distributions instead of mixture of Gaussians. Using our own proposed *sectorial extreme points* feature set, we have achieved the maximum possible classification accuracy (100%) over the Weizmann dataset. This achievement should be considered jointly with the fact that we have used a lightweight feature set.

On a different ground, and from the implementation viewpoint, surveillance software for automated human action recognition requires portability over a variety of platforms, from servers to mobile devices. The current products mainly target low level video analysis tasks, e.g. video annotation, instead of higher level ones, such as action recognition. Therefore, we explore the potential of the MPEG-7 standard to provide a standard interface platform (through descriptors and architectures) for human action recognition from surveillance cameras. As the last contribution of this work, we present two novel MPEG-7 descriptors, one symbolic and the other feature-based, alongside two different architectures: the *server-intensive* which is more suitable for "thin" client devices, such as PDAs and the *client-intensive* that is more appropriate for "thick" clients, such as desktops. We evaluate the proposed descriptors and architectures by way of a scenario analysis.

We believe that through the four contributions of this thesis, human action recognition systems have become more practical. While some contributions are specific to generative models such as the HMM, other contributions are more general and can be exploited with other classification approaches. We acknowledge that the entire area of human action recognition is progressing at an enormous pace, and that other outstanding issues are being resolved by research groups world-wide. We hope that the reader will enjoy the content of this work.